



Intitulé de l'offre de stage

Prédiction des hospitalisations de patients atteints de maladies respiratoires chroniques à partir de données environnementales

Stagiaire en	Informatique / Machine Learning
Affectation	CHU Toulouse
Durée	5 à 6 mois
Rémunération	environ 670 € net mensuel
Date de la publication	01/12/2024
Date d'embauche prévue	mars ou avril 2025 suivant disponibilités
Lieu	CHU, 2 rue Charles Viguerie, 31300 Toulouse, France

Le Centre Hospitalier Universitaire de Toulouse

Le Centre Hospitalier Universitaire (CHU) de Toulouse est constitué de plusieurs sites (les principaux étant les sites de Rangueil, Larrey et Purpan, ainsi que celui de l'oncopole en commun avec l'institut Claudius Régaud); il comprend 4 000 médecins et 12 000 personnels hospitaliers. La mission de recherche et d'innovation du CHU fait partie intégrante à la fois de son activité quotidienne et de sa stratégie pour l'avenir. Elle est menée en collaboration étroite avec les facultés et les organismes de recherche que sont notamment l'Institut national de la santé et de la recherche médicale (Inserm) et le Centre national de la recherche scientifique (CNRS).

L'Institut de Recherche en Informatique de Toulouse

L'Institut de Recherche en Informatique de Toulouse (IRIT), une des plus imposantes Unité Mixte de Recherche au niveau national, est l'un des piliers de la recherche en Occitanie avec ses 700 membres, permanents et non-permanents. De par son caractère multi-tutelle (CNRS, Universités toulousaines), son impact scientifique et ses interactions avec les autres domaines, le laboratoire constitue une des forces structurantes du paysage de l'informatique et de ses applications dans le monde du numérique, tant au niveau régional que national.

L'équipe d'accueil

L'informatique, le traitement des données et l'intelligence artificielle sont appelés à prendre une place croissante dans le monde de la recherche médicale. Dans ce cadre, le CHU de Toulouse a créé une équipe de "gestion de la donnée" dont le but est le recueil, l'analyse et le traitement de la donnée médicale, en support aux activités médicales du CHU. Elle collabore activement avec l'IRIT afin d'appliquer les techniques les plus récentes de la science des données et de l'intelligence artificielle dans le domaine de la santé.

Le stagiaire pourra également être amené à travailler avec le Pr Laurent Guilleminault au sein du service de pneumologie-allergologie à l'hôpital Larrey.

Objet du stage

L'épidémie de COVID19 a mis en lumière les difficultés pour les services hospitaliers à anticiper des pics soudains d'hospitalisations, menant à des délais dans l'adaptation de leur capacité d'accueil. Ces délais impliquent alors un fonctionnement en sous-effectif, ce qui provoque à la fois une prise en charge dégradée des patients et des conditions de travail plus difficiles pour les personnels médico-soignants.

L'objectif de ce stage est donc de prédire ces pics d'hospitalisation plusieurs jours à l'avance à partir de données environnementales et épidémiologiques pertinentes. Le sujet se restreindra aux hospitalisations causées par une décompensation de BronchoPneumopathie Chronique Obstructive (BPCO) ou d'asthme.

Démarche :

Le projet comporte deux grandes étapes :

- > La collecte des données pertinentes pour l'étude.
- > L'entraînement de modèles prédictifs de Machine Learning (ML) à partir des données collectées.

Collecte des Données :

Les données nécessaires à l'étude se séparent en deux catégories :

- > **Les données hospitalières** : ces données proviennent de différentes bases de données du CHU et doivent permettre d'identifier les patients à inclure dans l'étude, leurs date et motif d'hospitalisation ainsi que toute donnée clinique jugée pertinente par l'encadrement médical. Certaines de ces données pourront avoir été pré-extraites par le service des données du CHU afin d'accélérer cette étape du projet. Les données manquantes devront être extraites par la stagiaire en coordination avec son encadrant au sein du service des données.

- > **Les données environnementales** : ces données constitueront la majorité des variables d'entrée des modèles de ML. Il faudra d'abord déterminer avec l'aide de l'équipe médicale quelles sont les variables potentiellement prédictives d'une hospitalisation pour décompensation de BPCO ou d'asthme. Certaines ont déjà été identifiées, comme la pollution atmosphérique, les pollens, les épidémies infectieuses (grippe, COVID, etc.) ou les conditions météorologiques. Ces données devront prendre la forme de séries temporelles afin de pouvoir utiliser les informations liées à l'évolution de ces variables. Par la suite, il s'agira d'identifier des bases de données publiques (ou pouvant être mises à disposition dans le cadre de l'étude) présentant les variables d'intérêt, de les collecter et de les rendre interopérables les unes avec les autres (notamment en adaptant leurs résolutions spatiale et temporelle).

Entraînement des modèles :

Une fois les données collectées prétraitées, le projet passera à la construction de modèles d'apprentissage pour prédire l'évolution du nombre d'hospitalisations à court terme :

- > **Entraînement des modèles** : Utilisation de modèles de type LSTM (Long Short-Term Memory), CNN (Convolutional Neural Networks) et architectures combinées pour capturer les relations temporelles et structurelles des séries temporelles environnementales.
- > **Évaluation des performances** : Analyse de la précision prédictive des modèles via des métriques adaptées telles que l'AUROC, la calibration des prédictions, et l'analyse des matrices de confusion.

Environnement Technique :

Des données hospitalières seront extraites à l'aide de requêtes SQL. Les bibliothèques Python (Pandas, Polars, TensorFlow, PyTorch, Scikit-Learn) seront utilisées pour prétraiter les données, développer les modèles et implémenter les algorithmes de pré-traitement. Les séries temporelles seront traitées avec des frameworks dédiés comme TSFEL pour l'extraction des caractéristiques, ainsi que des architectures LSTM et CNN pour la modélisation.

Impact et Perspectives :

Ce projet vise à améliorer la capacité d'accueil des services hospitaliers lors des pics d'hospitalisation. En effet, la prédiction à court terme de ces pics doit permettre aux services de les anticiper et de mobiliser le personnel nécessaire pour la prise en charge des patients. Ce stage doit permettre de tirer profit des connaissances médicales actuelles et des données existantes afin de créer un outil concret d'aide à la décision pour les services hospitaliers.

Encadrement :

L'encadrant principal de ce stage sera Robin Schwov, data scientist au CHU de Toulouse et actuellement en thèse d'université sur la thématique de la prédiction du pronostic en oncologie par ML.

Formation

École d'ingénieur, de préférence avec spécialisation en informatique ou mathématiques (en année de césure ou stage long). Master 2 informatique ou mathématiques appliquées.

Compétences attendues

Des connaissances en science des données, apprentissage et réseaux de neurones seront appréciées.

Modalité de candidature

- > CV à envoyer à: pneumo@stages-medecine-numerique.fr
- > Date limite de candidature : 01/03/2025
- > Encadrants: Robin Schwob (CHU Toulouse), Pr Laurent Guilleminault (CHU Toulouse), Sylvain Cussat-Blanc (IRIT), Sandrine Mouysset (IRIT)

Une première sélection sera effectuée sur la base des CVs reçus. Les candidats seront immédiatement informés du résultat, et ceux qui seront retenus à l'issue de la première sélection seront invités à un ou deux entretiens individuels en téléconférence avant sélection définitive.

L'ensemble des offres de stage est disponible sur <http://www.stages-medecine-numerique.fr>.